Synergy of Database Techniques and Machine Learning Models for String Similarity Search and Join

Lu, Jiaheng

ACM
2019-11

Lu , J , Lin , C , Wang , J & Li , C 2019 , Synergy of Database Techniques and Machine Learning Models for String Similarity Search and Join . in CIKM '19 : Proceedings of the 28th ACM International Conference on Information and Knowledge Management . ACM , New York, NY , pp. 2975-2976 , ACM International Conference on Information and Knowledge Management , Beijing , China , 03/11/2019 . https://doi.org/10.1145/3357384.3360319

http://hdl.handle.net/10138/307683
https://doi.org/10.1145/3357384.3360319

unspecified
publishedVersion

Synergy of Database Techniques and Machine Learning Models for String Similarity Search and Join

Jiaheng Lu
University of Helsinki
jiaheng.lu@helsinki.fi

Chunbin Lin
Amazon AWS
lichunbi@amazon.com

Jin Wang
University of California, Los Angeles
jinwang@cs.ucla.edu

Chen Li
University of California Irvine
chenli@ics.uci.edu

ABSTRACT

String data is ubiquitous and string similarity search and join are critical to the applications of information retrieval, data integration, data cleaning, and also big data analytics. To support these operations, many techniques in the database and machine learning areas have been proposed independently. More precisely, in the database research area, there are techniques based on the filtering-and-verification framework that can not only achieve a high performance, but also provide guaranteed quality of results for given similarity functions. In the machine learning research area, string similarity processing is modeled as a problem of identifying similar text records; Specifically, the deep learning approaches use embedding techniques that map text to a low-dimensional continuous vector space.

In this tutorial, we review a number of studies of string similarity search and join in these two research areas. We divide the studies in each area into different categories. For each category, we provide a comprehensive review of the relevant works, and present the details of these solutions. We conclude this tutorial by pinpointing promising directions for future work to combine techniques in these two areas.

CCS CONCEPTS

• **Information systems** → **Information integration**; *Data extraction and integration*; • **Computing methodologies** → *Rule learning*.

KEYWORDS

databases, machine learning, string similarity join, string similarity search, data integration

1 INTRODUCTION

Today's society is immersed in a wealth of text data, ranging from news articles, to social media, research literature, medical records, and corporate reports. Identifying data referring to same real-world entities is a core task of information retrieval, data integration, data cleansing and data mining when integrating data from multiple sources. This problem can be resolved with string similarity search and join [1, 8, 9, 13–16]. Data referring to the same real-world entity is usually represented in different formats due to the following scenarios: (i) data stored in different sources may contain typos and be inconsistent; and (ii) data stored in different sources use different representations. It is very challenging to identify matched strings correctly. In addition, when scaling this problem to big data volume, it brings extra performance challenge.

In this tutorial, we introduce the state-of-the art techniques in database and machine learning areas to solve the challenges by providing high-quality matching approaches and efficient algorithms for string similarity search and join [2–4, 7, 11]. We first present the existing works in database area and machine learning area separately, then we discuss the connection between the techniques in these two different areas. In database area, a plethora of works are proposed to design various index structures to improve the performance of similarity string search and join. In machine learning area, similarity string matching is modeled as the problem of entity matching, which aims at identifying whether two entities are with the same identification. And different kinds of models are trained to find such entity pairs. Recently, deep learning techniques have been extensively adopted in identifying string semantic similarity, where texts are mapped into low-dimensional continuous vector space. They use embedding techniques to find matched entities. The building blocks of deep learning for string similarity measurement are mainly Recurrent Neural Network (RNN) [5] and distributed representation learning [10]. Figure 1 provides an overview of the history of the string similarity measures and entity matching techniques.

To the best of our knowledge, this is the first tutorial to discuss the synergy between database techniques and machine learning approaches on string similarity search and join. Note that the problem of string similarity matching was well studied in computer science. The first references to this problem could be traced to the sixties and seventies [12], where the problem appeared in a number of different fields, such as computational biology, signal processing, and text retrieval. However, this tutorial will review this problem from a novel angle through the synergy between database and machine learning techniques. The existing tutorial and survey review this problem separately from their respective fields. We have identified few tutorials (e.g. [6]) on string similarity search, which are mainly from VLDB and ICDE. However, they were given in 2009, ten years ago. This tutorial, on the other hand, focuses on the state-of-the-art

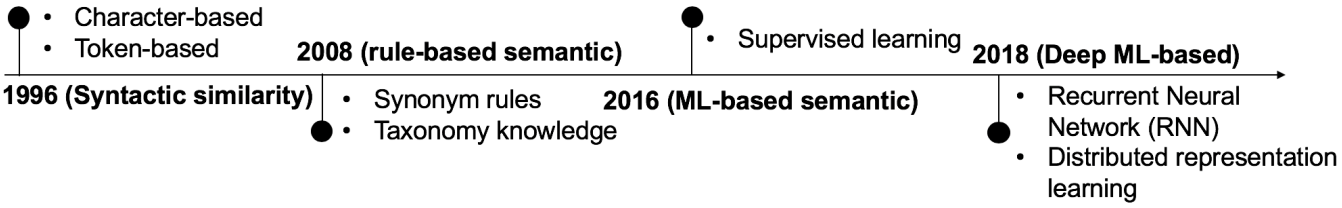

Figure 1: Recent 20 years of string similarity measures

works on string similarity search and join bridging the two fields of database and machine learning.

2 TUTORIAL ORGANIZATION

This tutorial consists of 4 parts and is planned for 3 hours. The details are explained as following.

Motivation and Background (30 minutes).

- A brief overview of the history of string similarity search and join.
- The real world applications of string similarity search and join.
- The formal problem definition and necessary background.

Database Related Techniques (60 Minutes).

- State-of-the-art algorithms about String similarity join.
- String similarity search algorithms.
- Enhancing string similarity and join with semantic features including synonym and taxonomy knowledge
- Application of string search and join techniques, including approximate entity extraction, query autocomplete and conjunction with other kinds of data, e.g. graph, spatial and streaming.

Machine Learning Related Techniques (60 Minutes).

- Traditional machine learning approaches for string similarity measurement.
- Preliminary about deep learning and its application in natural language processing.
- String similarity measurement with deep learning techniques.

Synergy between Database and Machine Learning (15 Minutes).

- Applying database filtering technique to enhance machine leaning algorithms.
- Incorporating machine leaning models in string similarity functions for string joins in databases.

Open Problems (15 Minutes).

- General purposed pipeline for string similarity search and join.
- Accelerating the machine learning based approaches.
- Combining Human-in-the-loop with machine learning approaches for better performance.

This tutorial can help not only motivated researchers and developers to select new topics and contribute their expertise on string similarity search and join, but also new developers and students to quickly build a comprehensive overview and grasp the latest trends and state-of-the-art techniques in this field.

3 SHORT BIBLIOGRAPHIES OF TUTORS

Jiaheng Lu is an Associate Professor at the University of Helsinki, Finland. His main research interests lie in the big data management and database systems, and specifically in the challenge of efficient data processing from real-life, massive data repository and Web.

Chunbin Lin is a software development engineer at Amazon Web Services (AWS) and he is working on AWS Redshift. His research interests are database management and big data analytics.

Jin Wang is a fourth year PhD student at the University of California, Los Angeles. His research interest mainly lies in the field of data management and text analytics.

Chen Li is a professor in the Department of Computer Science at UC Irvine. His research interests are in the field of data management, including data-intensive computing, query processing and optimization, visualization, and text analytics.

REFERENCES

[1] Roberto J. Bayardo, Yiming Ma, and Ramakrishnan Srikant. 2007. Scaling up all pairs similarity search. In *WWW*. 131–140.
[2] Alexander Behm, Shengyue Ji, Chen Li, and Jiaheng Lu. 2009. Space-Constrained Gram-Based Indexing for Efficient Approximate String Search. In *ICDE*. 604–615.
[3] Paul Suganthan G. C., Adel Ardalan, AnHai Doan, and Aditya Akella. 2018. Smurf: Self-Service String Matching Using Random Forests. *PVLDB* 12, 3 (2018), 278–291.
[4] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq R. Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed Representations of Tuples for Entity Resolution. *PVLDB* 11, 11 (2018), 1454–1467.
[5] Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science* 14, 2 (1990), 179–211.
[6] Marios Hadjieleftheriou and Chen Li. 2009. Efficient Approximate Search on String Collections. *PVLDB* 2, 2 (2009), 1660–1661.
[7] Chen Li, Jiaheng Lu, and Yiming Lu. 2008. Efficient Merging and Filtering Algorithms for Approximate String Searches. In *ICDE*. 257–266.
[8] Jiaheng Lu, Chunbin Lin, Wei Wang, Chen Li, and Haiyong Wang. 2013. String similarity measures and joins with synonyms. In *ACM SIGMOD*. 373–384.
[9] Jiaheng Lu, Chunbin Lin, Wei Wang, Chen Li, and Xiaokui Xiao. 2015. Boosting the Quality of Approximate String Matching by Synonyms. *ACM Trans. Database Syst.* 40, 3 (2015), 15:1–15:42.
[10] Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*. 234–239.
[11] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *SIGMOD*. 19–34.
[12] Gonzalo Navarro. 2001. A Guided Tour to Approximate String Matching. *ACM Comput. Surv.* 33, 1 (March 2001), 31–88.
[13] Chuan Xiao, Wei Wang, and Xuemin Lin. 2008. Ed-Join: an efficient algorithm for similarity joins with edit distance constraints. *PVLDB* 1, 1 (2008), 933–944.
[14] Pengfei Xu and Jiaheng Lu. 2017. Top-k String Auto-Completion with Synonyms. In *DASFAA*. 202–218.
[15] Pengfei Xu and Jiaheng Lu. 2018. Efficient Taxonomic Similarity Joins with Adaptive Overlap Constraint. In *ACM CIKM*. 1563–1566.
[16] Pengfei Xu and Jiaheng Lu. 2019. Towards a Unified Framework for String Similarity Joins. *PVLDB* 12, 11 (2019), 1289–1302.